DOCUMENT RESUME

ED 307 896 IR 052 785

AUTHOR Triebwasser, Marc A.
TITLE The Electronic Library: The Student/Scholar Workstation, CD-ROM and Hypertext.
PUB DATE 88
NOTE 30p.; Paper presented at the Annual Meeting of the American Political Science Association (Washington, DC, August 31-September 4, 1988).
PUB TYPE Information Analyses (070) -- Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Access to Information; Character Recognition; Computer Networks; Copyrights; *Databases; *Electronic Publishing; *Information Retrieval; *Information Storage; Microcomputers; *Optical Data Disks; Optical Scanners; Relevance (Information Retrieval); Standards; Telecommunications
IDENTIFIERS Hypermedia; Hypertext; Workstations

ABSTRACT

Predicting that a large component of the library of the not so distant future will be an electronic network of file servers where information is stored for access by personal computer workstations in remote locations as well as the library, this paper discusses innovative computer technologies--particularly CD-ROM (Compact Disk-Read Only Memory) and Hypertext--and their roles in the electronic library. The major topics presented include: (1) the development of powerful microprocessors and the capacity of this technology to provide analogue representation; (2) the new trend of information in databases; (3) hypertext (the ability to link information in one portion of text with other information in another place in that same document or in another document) and hypermedia (the linkage between a piece of text and information presented in nontextual forms); (4) the functional importance of the asymmetric technology of CD-ROMs; (5) storage needs of different types of information; (6) how machine readable text is stored on CD-ROM; (7) optical character recognition technologies (OCR); (8) advantages and disadvantages of electronic imaging; (9) technological impediments (OCR technology, high resolution monitors, fast networks); and (10) institutional impediments (copyright and economic considerations, lack of standards for CD-ROMs). Concluding remarks discuss the possibility of commercial markets for information on CD-ROM; the cost effectiveness of making academic journals available on CD-ROM; and projections for future use of this technology. (CGD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307896

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

THE ELECTRONIC LIBRARY:
The Student/Scholar Workstation, CD-ROM and Hypertext

Marc A. Triebwasser
Central Connecticut
State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Marc A. Triebwasser

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IR053185

Prepared for delivery at the 1988 Annual Meeting of the American Political Science Association, The Washington Hilton, Washington, D.C., August 31 - September 4, 1988.

Not for attribution or quotation without the permission of the author. (c) Marc A. Triebwasser, 1988

This paper represents work in progress and does not contain the footnotes and bibliographic references which will appear in the final version. Please contact the author for permission for attribution or quotation:

Dr. Marc A. Triebwasser
Department of Political Science
Central Connecticut State University
New Britain, Connecticut 06050

(203) 827-7328.

THE ELECTRONIC LIBRARY:
The Student/Scholar Workstation, CD-ROM and Hypertext

Marc A. Triebwasser
Central Connecticut
State University

Computers in political science research and teaching have for some time been the purview of behavioralists and others interested in quantitative methodologies. However, with the advent of very powerful microcomputers and such technologies as CD-ROM (Compact Disk-Read Only Memory)--through which an entire encyclopedia can be stored on a single 5-1/4" compact disk--computers are now in a position to serve those in the social sciences and humanities who are involved in more traditional, library oriented research.

What makes these new computer technologies so appropriate to traditional research is the fact that we now have a method not only of storing vast quantities of information, but also of including with this information methods for accessing it quickly and intelligently. It is the availability of increasingly sophisticated information retrieval techniques that make CD-ROM and similar storage media so different from all previously available methods of information storage. It is simply not enough to have a great deal of information; one must have easy and intelligent access to the information if it is to be truly useful.

In fact, we are moving toward a totally new concept, that of an *electronic library* in which information will be stored in computer readable form on optical, magnetic or other media, and will be available to the user on

(c) Marc A. Triebwasser, 1988

personal computer workstations located in the library, in scholars' offices, in students' dorm rooms, and even at home. A large component of the library of the not so distant future will thus be an electronic network of file servers where information is stored, and personal workstations where information is accessed.

Analogue Representation

This is all made possible by the development of powerful microprocessors (the Intel 80386 (386) processor, used in the IBM compatible world; and the Motorola 68020 and 68030, used in the Apple Macintosh II), high resolution monitors, high speed networks, and high capacity storage devices. These technologies enable the computer to report things in analogue and graphics form, the way human beings think; rather than in digital form, the way today's computers process information. This difference can be seen, for example, by comparing a traditional analogue clock face with a more recent digital one. Both clocks will tell you that it is, say, 4:13 in the afternoon. However, the relationship between the minute hand and the face of the analogue clock immediately makes one aware that it is almost a quarter past the hour. And the relationship between the hour hand and face tells one that it is about two-thirds of the way through the afternoon. The process of describing a line presents us with another example of analogue thinking. It is much easier for human beings to deal with a graphics representation of a line, rather than being given, or having to give, the coordinates of its end of points. Similarly, it is usually easier for human beings to grasp data if it is represented graphically, rather than being presented as a table of numbers.

This way of thinking about how information could most usefully be presented to human beings is generally

thought to have begun with an article entitled "As We May Think," which Vannevar Bush published in *The Atlantic Monthly* as early as 1945. And much research in the area of information presentation and the machine-human interface was carried out in the seventies in such institutions as Xerox's Palo Alto Research Center (Xerox PARC) and the Stanford Research Institute (SRI). During this time period, Xerox developed a computer--called the Star--which used a graphics interface based upon the research which had been going on at PARC, and this technology became the basis for the interface which Apple now uses on its Macintosh computers.

The reason for Apple calling the Macintosh "the computer for the rest of us" is that it presents information in a graphic analogue format. Thus, it uses the power of the computer to translate information into forms readily accessible to human beings; rather than forcing human beings to learn command structures and digital methods of representing instructions and information, the way machines work. Through such interfaces, instead of having to memorize commands to instruct the computer to perform a certain operation, one simply makes choices from *menus* of possibilities. Often the options are represented as graphics figures or *icons*. This technology also allows the user to have different *windows* or areas of the screen dedicated to different tasks, so that the user may quickly move between one operation or another and compare the results of several different operations. Furthermore, choices are made by manipulating a *pointing device*--usually a mouse--rather than using the keyboard.

This approach to a user interface has now entered the IBM compatible world through Microsoft's Windows and IBM's Presentation Manager. Through the use of such techniques, we are approaching the situation in

which the machine can do most of the drudgery work in searching and accessing information, and can present this information in forms easily assimilated by human beings. People can then spend most of their time in the more creative work of analyzing and juxtaposing this information to arrive at new ideas and conceptualizations.

Types of Databases

In the past, computers were mainly used to deal with numeric information. However, this is changing rapidly. Most scholars are, for example, familiar with word processors and the ability of the computers to manipulate full text. We are now moving from a time when computer databases usually consisted of numeric, character, date, and logical fields to a time when they also deal with fields containing many paragraphs of text, and even graphics, music, and photographic images. And we are quickly approaching the ability to handle full motion digital video as well.

Many libraries have already used the computer to automate their acquisition, cataloguing and circulation processes. Many have also used national services in order to access bibliographic information. Up until now, these services have provided access to online databases which are stored on large centralized computers and accessed over the telephone. Several problems are encountered with the use of these databases, however. First of all, the use of these databases is rather expensive. The costs involved include annual membership fees, online connect time charges, and telephone charges as well. For this reason, at many libraries these services cannot be accessed directly by the end users--be they students or faculty. They are rather accessed by a member of the library staff who is very familiar with the way these various databases work

and have acquired the skill to develop effective search strategies, so as to minimize the actual time online and thus reduce connect time and telephone costs.

Another problem encountered with the use of these databases is that once one has located the reference works one wants to use, the local library often does not have the journal or book on hand. These must therefore be ordered through interlibrary loan, a process which often takes several weeks. This often proves very discouraging for scholars, and especially for students who often need to get their research finished and their term papers handed in *yesterday*.

Through the use of CD-ROMs, computer tapes and even floppy disks, bibliographic information can now be published electronically and distributed so that they can be put online at the local library or research site. This eliminates connect time and telecommunication charges, and gives libraries the ability to allow end users to conduct their own searches, and thus to acquire the skills necessary to develop competent research strategies. With the advent of as much information as is available today, it is impossible for us to teach students even a small fraction of what they might need or like to know. What we need to do is to teach them how to find relevant information and how to use it wisely. Unless they have the opportunity to conduct database searches themselves, we are not providing our students with the skills they need.

Full Text Databases

Some national bibliographic database vendors have now developed programs to distribute some of their databases, either on computer tapes or on CD-ROMs, so they can be mounted and searched on *local* computers. Such a service is now provided, for example, by BRS/OnSite. The ability of CD-ROM to store vast

quantities of information means that we now also have the ability to distribute *full text* information. This is particularly important because it means we will eventually be able to distribute current and retrospective issues of academic journals, magazines, and newspapers in computer readable form on compact disks, in such a way that the information contained on them can be accessed easily and intelligently. In the future, one will therefore not only have the ability to discover information sources which are likely to contain information relevant to one's needs, but will also be able to have immediate access to the articles, books, and other informational sources referred to in the bibliographic databases. Moreover, through electronic imaging techniques, one can easily have access to photograph-like copies of documents stored on CD-ROMs. A wide variety of researchers would thus have access to electronic copies of rare books and documents--access to which is now limited to only a select few.

Hypertext and Hypermedia

Another advantage of having information in computer readable form is the ability to link information in one portion of text with other information in another place in that same document or in another document. This technique is called *hypertext*. Although this term is being used a great deal today in a high tech environment, the actual concept has been with us for some time. In fact, we have all used a hypertext device to link information located in different places, probably without realizing we were doing so. This familiar hypertext device is the *footnote*.

However, the application of this device in the manual and electronic worlds is quite different. Traditionally, when coming upon a footnote one would have to spend

a considerable amount of time locating the reference work to which the link is being made and then spend more time finding the actual material. In the electronic world, this could be done almost instantaneously if the material is available in computer readable form, and the proper linkage mechanisms have been put into place.

Hypertext concepts in computer readable material can be implemented rather easily. A word which is linked to other pieces of information can be highlighted in some way, or a symbol could be used in the on-screen text indicating that a link is present. Then simply by pressing a button or two, the place with which this link or pointer connects can be brought onto the screen, either replacing the current text or in another window, so that the current text can be compared directly with the text to which it is linked. Hypertext products require a great deal of work to develop; and work in this area has been carried out at several universities, including Brown. At Brown University, a Hypermedia Project has been under way for several years, first being used to teach two courses and now being applied to a number of additional ones.

The name of this project brings into focus another dimension of textual linkage. Not only can links be made between one piece of text and another, links can also be made between a piece of text and information presented in nontextual forms, such as graphics, photographs, voice, music, sound, simulations, and even full motion video--all of which can be stored in a digital format. Thus, in addition to hypertext, we also have the possibilities of *hypermedia* links between one medium of information and another.

Up until now, such hypermedia linking has been accomplished by having a computer connected to an audio tape recorder, video disk player, microfiche or microfilm reader, or similar devices and having the

computer direct the machine to present the desired externally stored information. With the advent of CD-ROM technology, information which has heretofore required storage on different media can now all be stored in digital or analogue form on the same optical medium as the software. What this means is that what is now being attached to CD-ROM drives is not only computers, but also video screens, sound boards, speakers, and the like.

The possibilities with which these technologies present us are awe inspiring indeed. In reading an electronic text about Beethoven, one could call up a recording of his Ninth Symphony, for example; or in reading about John F. Kennedy, one could call up a recording of his inaugural address, allowing the user to hear the speech spoken in Kennedy's own voice. In reading about military conflicts, one could have access to maps onscreen on which the progress of battles could be simulated through animation. And in reading about more recent political events, one can have access to television footage actually showing the event as it took place. These capabilities will fundamentally change the way we think about information, and the way we store and access it. It will allow us to move beyond the one dimension of simple text and make traditional library research infinitely richer and more interesting.

Through hypertext and hypermedia techniques, the student or scholar will no longer be limited to reading through a book or document in a linear fashion, but will be able to navigate through the book as he or she wishes using such conceptual links as common theme, and will also be able to jump easily from one work to another and from one medium to another. Individuals will thus be able to navigate through humankind's vast store of information the way they want, rather than in ways

imposed on them by outside authorities or by the limitations of older technologies.

Read Only: An Asymmetric Technology

CD-ROM represents a significant technology, not only because of the capabilities of the medium itself, but because of the way one stores and retrieves information on it. It has thus lead to the development of new conceptualizations of how information is stored and accessed. It is the *functionality* of these disks, rather than the *medium* itself, which perhaps represents the greatest advance. And even if CD-ROMs themselves are replaced by other media in the future, this functionality will be with us for some time.

One of the factors in the functional importance of CD-ROMs is that being a read only medium, its use is *asymmetric*. More traditional computer media, such as disks and tapes, could be written to, read from, erased and rewritten. This meant that attention had to be paid both to the process of storing information on these media and to that of retrieving information from them, and both processes needed to be optimized at the same time. With a read only medium, one can develop mechanisms where information can be accessed very quickly and intelligently, and not have to worry how long it takes to process information for storage on these disks. And here we are not talking about the actual electronic process of getting information from these disks, but rather the process of *locating information* most relevant to one's needs.

Full text CD-ROMS usually contain more information than simply the text itself. Often half of the disk space used is devoted to text, and the other half to storing an *inverted index* in which all uses of non-noise words (such as "a," "the," "of," etc.) are listed. In more traditional terms, this inverted index would be called a

concordance. As other retrieval techniques are developed, other types of information locating devices will also be stored on these disks. The place on the disk where the text is stored is also important, so that in some cases schemes are developed to place related textual and other information in close physical proximity to each other.

Storage Needs of Different Types of Information

As we have noted, many different types of information can be stored on a CD-ROM, including numerical data, full text, computer programs, simulations, graphics, photographs, and audio and video information. However, these types of information require different amounts of storage capacity. Thus, for example, a compact disk can currently hold about 200,000 pages of printed textual information or about an hour of high fidelity music, but only ten minutes of video--and, using current techniques, that video would be stored in an analogue, rather than a digital, format.

The amount of information needed to store different forms of information is an important factor, as can be seen in the method of the representing of text on a computer screen. If information is stored in a *character* format, most common computer screens will hold 80 characters per line and 25 lines per screen or a total of 2,000 pieces of information per screen. However, if the same information is stored as an actual *image* (or electronic photograph) of the page, then much more information is needed.

At the lowest resolution used on computer screens, we would have 320 dots on each horizontal line and 200 such lines per screen. Each such dot is called a pixel, or *pic*ture *el*ement. This means that such a representation would necessitate 64,000 pieces of information. In addition to this, if we are dealing in color, then we have

to know how many different colors or shades we want to include. The larger number of colors or shades to be represented, the greater amount of information needed for each pixel. The same amount of information which is used to represent characters would allow us to represent 256 different shades.

This is usually quite good for graphics, but not necessarily for a photograph. In a photographic representation of a human face alone, there are well over 256 shades of color. For truly high quality color, one really needs to store three times the amount of information needed to represent a character at each of the 64,000 locations or pixel on even a low resolution screen. Of course, if one were representing color that carefully, one would also be using a screen with a higher resolution, with something like a million picture elements per screen. If we use the equivalent of three characters (or bytes) of information per picture element and we are using one million picture elements per screen, we are talking about a great deal of information needed to represent a single screen--some *three million* pieces of information.

Fortunately, we are able to reduce this amount of information somewhat by using various technologies through which this information can be compressed into a smaller amount. Nevertheless, representing information in a graphics format takes far more space than representing that same information in character format. Moreover, when one talks about full motion video and takes into account the fact that the common standard for video representation is to use 30 frames a second--because the human eye interprets this as continuous motion--one recognizes the vast amount of information needed to represent video in a digital format.

Machine Readable Text

As we move into the area of putting the full text of articles, books, and the like onto CD-ROMs, there are a number of matters which must be considered. One of them is how do we get the text into a format which can be stored on CD-ROMs. One approach is that of translating the information contained in the articles into computer readable form by representing the information as *individual characters*. If one has to rekey in all of this information from its printed form, the process can be quite costly, even when the actual work is done off shore. In addition to this, as part of this process one actually needs to have this information rekeyed two or three times for verification purposes.

One would hope that for the more recent materials at least, one could avoid rekeying the textual information by making use of the information already in computer readable form on the magnetic tapes used for typesetting purposes in publishing books and journals today. These tapes, however, contain much typesetting code which must be eliminated before the text can be put onto CD-ROMs. One of the special difficulties involved here is that these codes are sometimes intermingled with the individual words themselves as, for example, when hyphenation occurs. In addition to this, a great many different coding schemes are employed by the various makes and types of typesetting machines used today.

To deal with this, the publishing industry has attempted to standardize this coding. This new approach is called the *Standard Generic Markup Language* (SGML). When this comes into wide use, it will be rather easy to translate typesetting tapes into information to be put on CD-ROMs. For the time being, however, we have to deal with the many different typesetting codes currently in use, and must develop software which can translate text containing each of

these coding schemes into a form appropriate for CD-ROM. In the long run, this approach may prove less expensive than rekeying the information.

One of the advantages of SGML is that it identifies different parts of text as different types of *functional objects*. An actual document contains many different types of functional objects. These include section headings, subsection headings, and the like; main body paragraphs, indented quotes, and footnotes; bibliographic material, lists and tables; etc. Up until now, for most typesetting purposes, these various elements have been identified by specifying particular type faces and sizes, margin and tab settings, line placement, and so forth. Identifying these elements or objects in this way makes it difficult to translate information from one medium to another. What SGML tries to do is to identify each textual object functionally by labeling it as a chapter title, subsection heading, main body paragraph, footnote, etc.

This information is then combined with a *style sheet* containing information on how to handle these various textual elements. The style sheet may suggest, for example, that subsection headings should be printed in 12 point Bodoni bold, in upper and lower case, and placed left flush on the page with so much space above it and below it. A style sheet for placing the information on a CD-ROM would, of course, differ significantly from one used for typesetting purposes, but the functionally coded text used in conjunction with it would be exactly the same. The software for placing the text on a CD-ROM would then be able to combine the text and textual object information with the CD-ROM's style sheet and produce the appropriate input for placement on a CD-ROM. It is this *functional tagging* of textual objects which gives a generic, and therefore

transportable, quality to text annotated with the Standard Generic Markup Language.

OCR Technologies

Another approach to handling textual material is to scan it graphically with an optical reader and translate it into a digitized format. This electronic information can be interpreted by appropriate software using *optical character recognition* (OCR) technologies. Through these techniques, the text can be translated into individual characters and thus into a computer readable format. While it is fairly easy to do this for typewritten and dot matrix printouts where each letter is the same width, it becomes more difficult to scan typeset material where each letter is proportionately spaced. And the introduction of various type faces and sizes with bolding, italicizing, and the like makes the OCR process even more difficult. Moreover, even when such techniques are successful, one most often loses information about titles, section headings, tables, and so forth.

Because the OCR process is often a complicated and time-consuming one, the scanning of each page may take a significant amount of time. Recently, software has been developed which can interpret characters after they have been scanned and converted into an electronic image. The advantage to this approach is that the mechanical process of scanning each page of the document can be done rather quickly, and the computer can then later take its time in translating this imaged information into individual characters. Since the latter process can be done automatically, without human or mechanical intervention, this technique increases the speed of the actual scanning process and can cut costs considerably.

These approaches do advance OCR technology, but OCR still presents us with problems. Even if the

process is 99% accurate, it still means one has a misreading of a character once every one or two lines of text. Where such error rates are unacceptable, correcting these mistakes requires human intervention, and--in many cases--simply rekeying the information may be more cost effective.

Electronic Imaging

Another approach is to scan a page and simply create an *electronic image* of that page. This process is called electronic imaging and is similar to the process of microfilming documents. One of the advantages of this method is that one retains all the layout information on the page, including diagrams, tables, photographs, etc. Of course, if one wants to maintain high quality graphic information, one has to scan pages with considerably high resolution and with an adequate allowance for the various shades of gray. Some of the disadvantages of imaging is that the material is not available in a computer readable format, and therefore *cannot be searched*. Nor can it be blocked and selected for transfer to a word processing program, as can information in computer readable form. In addition to this, one runs into the problem discussed before that material represented in an imaged format requires more storage space and processing time than would the same material represented in character format.

One approach to solving this problem would be to image the document and use OCR technology to get it into character format, even with some errors. One would then use the character version to create an inverted index and place that inverted index along with the *imaged version* of the document, but not the OCR version, onto the CD-ROM. One could then use the inverted index for search purposes, and present on screen the imaged versions of the page identified by the

search. There are disadvantages to this in that some occurrences of the words upon which the search is conducted may be missed due to misspellings introduced in the OCR process. Moreover, since the computer would not be able to recognize the individual words on the imaged page, the search words could not be highlighted. Nevertheless, this technique might prove helpful for getting a great deal of retrospective material onto a CD-ROM quickly, economically, and in a reasonably useful format.

Technological Impediments

OCR Technology. One of the technological impediments to getting full text scholarly information onto CD-ROMs is the need for improving OCR technology. As we have seen, even a 1% error rate can cause a significant need for human intervention and editing. As OCR technologies are improved, one would hope that human intervention would become unnecessary in the scanning process, and that this process could thus be fully automated. The development of more intelligent software that would correct misreadings would also prove very helpful. However, this latter approach would require a great many advances in the ability of computers to syntactically and semantically analyze text--an area we are just beginning to explore.

High Resolution Monitors. Another technological area in which improvement is needed is that of high resolution monitors. At the present time, the IBM compatible world is generally moving from an Enhanced Graphics Adapter (EGA) standard of 640 dots by 350 lines per screen to a VGA standard of 640 dots by 480 lines. However, much of the software currently on the market has not yet been updated to utilize the higher resolution of VGA monitors. Moreover, there is some

pressure to move the industry in the direction of monitors with 1,280 dots by 1,024 lines of resolution. This will produce screen images of very high impact and clarity. However, simple multiplication indicates that we are now talking about screens necessitating the addressing of some 1.3 million picture elements. This requires the use of more powerful processors and a good deal of memory. Unfortunately, many of those in the more traditional fields of the social sciences and the humanities who currently use computers use them mostly--if not exclusively--for word processing purposes. They therefore usually use monitors with character, and not graphics, capabilities. There will thus exist a real need to move these and other academic uses to displays with reasonably high resolution graphics capabilities.

Fast Networks. A third area of technological improvement which will be especially necessitated by the use of imaged documents is that of data communications. Currently, most local area computer networks are able to handle data at the rate of 10 megabits per second. The exact meaning of "megabits per second" is not that important for our purposes here. What is important is that many suggest that for handling the type of graphics which imaged documents represent, we are likely to need networks with ten times that capacity or, in other words, networks that are able to handle 100 megabits per second. And for digitized full motion video, we are likely to need networks with one hundred times today's capacity--ones which are able to transmit *one billion bits per second*.

Institutional Impediments

Most, if not all, of the technological problems are likely to be solved reasonably quickly, and one would hope that through the widescale deployment of these improved technologies, their costs would decrease

substantially--something which has been the general pattern in high technology developments. Of more concern are some of the institutional impediments to the development of the electronic library and the placing of vast stores of humankind's knowledge in computer readable, or at least computer accessible, form.

Copyright Considerations. Electronic publishing might be grouped into two categories. The first is the publishing of material in a digital format and distributing it on magnetic, optical or other media. This aspect resembles traditional publishing to at least some extent. One is still providing the user with a single copy of one's work, only in a different and more powerful form. Copyright protection of this type of material may be handled with reasonable facility, even though it is far more easier to make copies of material in electronic format than it is to photocopy large amounts of printed materials.

The far more difficult problem involves the *telecommunication* of materials in electronic format. The capability of transmitting materials available in electronic form means that many copies of one source may be displayed on any number of screens, many of which may be at remote locations. This is an aspect of electronic publishing with which the owners of information are likely to be more concerned.

In attempting to deal with this problem, Clifford Lynch of the University of California has suggested making a distinction between the simple *display* of information, and its actual *reproduction* in an electronic or hard-copy format. The simple display of information might be handled in one way, and its duplication in another. A model that could be used for the display of information, advanced by Clifford Lynch, is that of performance. In music, permission for the performance of works is handled through clearing houses such as

ASCAP and BMI. These centralized clearing houses grant permission for the performance of musical works, and in return collect royalties which are later turned over to the owners of these works. If such a scheme does prove acceptable to publishers, much of this clearing house function could be handled easily through the use of appropriate software stored on these disks and set into operation in conjunction with the process of accessing information.

Ithiel de Sola Pool has gone further and suggested that the electronic telecommunication of information might be seen as being analogous to the transmission of information by word of mouth, and therefore not subject to copyright protection. It is doubtful, however, that this interpretation will be widely accepted.

Recently adopted copyright law indicates that in using copyrighted material, one can only display it on one device at a time and that device must be located in reasonable proximity to the source material. Obviously, new legal and contractual mechanisms need to be developed for the distribution and telecommunication of information in electronic format. Only in this way will users have reasonable access to this information and authors and publishers receive reasonable remuneration for their efforts and property. With the revolution in information distribution which is now occurring, the whole question of *intellectual property* becomes a very significant one for which new solutions must be devised on many different levels.

Economic Considerations. With the success of music CDs, the price of mastering and reproducing CD-ROMs has decreased significantly. Software for indexing materials and performing other premastering functions is also becoming more generally available at reasonable prices. There is also considerable competition among vendors of access software and, as

more use is made of this software, it is likely to both improve in quality and come down in price.

The real economic stumbling block in the publication of information on CD-ROMs has to do with the cost of the information itself, rather than with the electronic publication of it. Much of the material currently available on CDs comes from government sources and is available without charge, or is in some other way already available in the public domain. Thus, we have a number of versions of ERIC and MEDLINE available today. The other type of information which is becoming more readily available for general use on CD-ROMs is in areas where there is a great commercial market for them. It is for this reason that we see much of the current CD-ROM publications available in the financial area, in the legal arena, in technology and the hard sciences, and in medicine. We will also be seeing quite a bit of information about Congress available in CD-ROM format, since such information is highly useful to lobbyists.

A third type CD applications which have become popular is the in-house publication of CD-ROMs by companies and institutions who need to distribute a great deal of proprietary information throughout their organization. For the most part, these applications represent various *vertical markets*. There has not yet been a significant development of general audience CD-ROM products which would represent a *horizontal market*, and would be of most use to scholars and libraries. This may soon change, however. Just as the use of CDs for entertainment has brought down the cost of producing them for informational purposes, the more widespread use of optical media players may so increase the installed base of home equipment capable of servingas CD-ROM players as to make the publication

of general audience CD-ROMs more economically feasible.

For example, the commercial production of an optical medium player that can play music CDs, video disks of various sorts, and CD-ROMs as well, is now being considered. If such devices become popular, then economies of scale will ensue and significantly lower the price of such players. This, in turn, will increase the installed base of CD-ROM players to a point where publishers will see a significant market for their products.

Standards. Another obstacle to the development of horizontal markets for CD-ROMs involves standards. The CD music industry provides a standard for the way information is physically placed on CDs. And the High Sierra format, as adopted by the International Standards Organization (ISO), provides a standard for file structures on CD-ROMs. Moreover, Microsoft, Sony, and Phillips have now promulgated a standard for the inclusion of audio and graphics on CD-ROMs. These standards ensure that the same CD-ROM can be played on CD-ROM readers from various manufacturers. The acceptance of this standard also ensures that a CD-ROM manufactured for IBM compatible computers can also be played on a reader manufactured for the Apple Macintosh, provided the appropriate software is included.

However, there are a number of other areas in which at least some degree of standardization needs to be achieved in order for users to be able to move easily from the use of one CD-ROM product to the use of another. These areas of standardization include the way text is represented on CD-ROMs, the way it is indexed, the mechanisms for accessing information on CDs (often called the *retrieval engine*), and the interface on the

computer screen which the end user accesses in order to obtain the information desired.

One of the greatest problems that librarians now have in the use of the various CD-ROM products currently available is that each CD-ROM database is accompanied with its own proprietary retrieval software and user interface. This means that librarians currently have to teach the user a different method for retrieving information with each database used. As more databases become available on CD-ROMs, such a proliferation of retrieval software and user interfaces will render impossible the efficient utilization of these databases for academic and library purposes. To achieve reasonable use of multiple CD-ROM databases--perhaps the very definition of a horizontal market--some degree of standardization must be achieved.

Standardization in the way a full text is represented on CD-ROMs is likely to be aided through the use of the Standard Generic Markup Language (SGML). However, better methods for the onscreen representation of such types of information as large tables need to be developed. And a similar problem exists with the onscreen representation of large charts and maps, in such a way as to allow for both detailed presentation and a sense of orientation of the detail to the whole.

Standardization in how the text is indexed and the retrieval software utilized to access it are interrelated matters. Different types of retrieval software often make use of specialized indexing methodologies and the inclusion of other special types of information along with the full text on the CD-ROM itself. For some time now, *Boolean strategies* have been utilized for accessing information in bibliographic, as well as full text, databases. These have included searching on the

occurrence of Word A *and* Word B in the same paragraph or field, or in a certain physical proximity to each other. It has also included searches for Word A *or* Word B; or Word A, but *not* Word B; and various combinations of these operations. Advances in this type of search strategy are now being developed.

One measure of a good search is the *relevance* of the material obtained. Another is the *completeness* of the information gathered. Unfortunately, methods designed to ensure the relevance of material often cause a less than complete listing of all related material. On the other hand, attempts to achieve completeness will often result in the listing of some sources which may not prove relevant. One of the methods used in some software to get around this dilemma is *priority ordering* of sources according to their likely relevance. A rather complete list of sources can thus be presented, and the user can choose just how far down the list he or she wishes to go.

Another approach is to attempt to identify *concepts* as clusters of words and to conduct searches based on these conceptual definitions. Yet another approach to better retrieval software is through the use of *expert systems*. In using such software, the user is asked a number of questions to help pinpoint exactly what information is desired, much in the same way a reference librarian may pose questions to a patron in order to hone in on the specific material required.

Retrieval software is thus becoming far more sophisticated as new approaches are developed. Moreover, different types of databases may require different types of software to optimize the access of information. For these reasons, there are those who argue that too soon an adoption of standards for retrieval software may prove dysfunctional in that it may prevent the development and widespread use of better approaches to information access. Nonetheless, the need

for the end user to be able to utilize a large number of different CD-ROM databases requires some degree of consistency between different CD-ROM databases.

Even when the retrieval software differs, some similarities in user interfaces may be possible. For example, certain *functional elements* of searches may be made similar. In the current evolution of microcomputer software, one has seen the development of some functional equivalence across different types of software--even in the IBM compatible world. For example, in many software packages, the F1 key is used to obtain on screen help, and the Escape key is used to move to the previous menu of choices, or to extricate oneself from a particular function. In the Macintosh world, a large degree of consistency has been fostered between different applications software packages, especially in their use of graphics interfaces. A similar consistency will hopefully also be developed in the use of Microsoft's Windows operating environment and the Presentation Manager graphics interface in OS/2s. Such commonalities allow users to learn new software packages more quickly. The inclusion of functional standards in CD-ROM database user interfaces would similarly allow the end user to move easily from the use of one database to that of another, even if different databases are enriched with particular individual functionalities.

One of the ways in which some standardization is beginning to take place in CD-ROM publication is through the publication of a number of different databases by a single publisher. Silver Platter was founded by people who are quite familiar with the needs of librarians, and this company has attempted to use a standardized user interface with all of the various database products it publishes. As other publishers begin to publish a number of different databases, some

similar standardizations may be achieved. Another possibility in the development of some standardization in user interfaces will likely come with the licensing of retrieval software to different publishers. As certain retrieval software packages become more popular, market forces may lead to the dominance of some particular user interfaces. But no matter how it comes about, in order for horizontal markets to develop for CD products, CD-ROM publishers are going to have to pay more attention to the needs of academic and public libraries--and that means to the development of at least some degree of standardization.

Conclusion

The CD-ROMs which are currently available are for the most part bibliographic in nature. In fields relating to political science, they cover such areas as business, the legal profession, and Congress, since these are areas of information for which there exist sizeable commercial markets. However, in order for CD-ROMs to be truly useful in the academy, the type of information available on CD-ROM will have to be expanded considerably.

As far as scholars are concerned, *academic journals* are one of the more important types of material to have available in computer readable form. However, their limited audiences make the publication of these journals on CD-ROMs a more difficult economic proposition. Some journal publishers are afraid that if they make their material available in an electronic format, the sales of their hard copy versions may erode. Since this is the bread and butter of such publishing houses--as well as many scholarly societies--this issue is approached with some concern.

There are, however, ways of handling these concerns. The *current year's issues* of a journal might be published at first *only* in hard copy format, as it is now, and then

be published in CD-ROM format *retrospectively*. This would allow scholars the ability to do electronic searches of all previous years' volumes and only have to do manual searches of the current year's information. This approach would thus safeguard the society's or publishing house's current earnings. In addition to this, CDs might be produced as a cooperative effort, containing many different journals. Individuals and libraries could then subscribe to specific journals, and the CDs could be coded or manufactured in such a way that the journals not subscribed to would be electronically or physically unavailable for access.

Since the availability of information in electronic format will make it possible to advance scholarship considerably, there are many national advantages to doing this. One might, therefore, also consider the possibility of *government grants* to provide seed money for these projects.

My general feeling, however, is that the availability of material in CD-ROM format will produce *large secondary markets* for these scholarly materials. Many libraries today have to limit their journal collections simply because of the lack of adequate shelf space. Obviously, distributing academic journals on CD-ROM vastly reduces this problem. Many journals are also of very specific interest and thus do not enjoy wide circulation. If small circulation journals are included on a CD-ROM with other journals, then that CD might be distributed to a far wider audience of scholars and libraries, and thus enjoy greater revenues.

At the present time, many libraries associated with state universities and liberal arts colleges cannot afford to maintain journal collections similar to those of the more prestigious research institutions. The publication of journals in CD-ROM format is likely to bring down the cost of such subscriptions and, by weighing the

advantages of volume versus those of margin, publishers may find that the increased availability of their journals to wider markets is likely to increase, rather than decrease, their revenues--even as they reduce the price of individual subscriptions to their journals in CD-ROM format. One consideration here is that it is often cheaper to distribute information on CD-ROMs than it is on paper. Since electronic publication is a relatively new possibility, it will take some time to convince publishers of these advantages, but it is certainly to the advantage of the scholarly community to press *strongly* for such a change in publication strategies.